DOCUMENT RESUME

ED 208 057 TM 810 820

AUTHOR Herman, David O.
TITLE Reducing Sex Bias in Ability Tests.
PUB DATE 24 Aug 81
NOTE 9p.; Paper presented at the Annual Meeting of the American Psychological Association (Los Angeles, CA, August, 1981).

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Academic Ability; *Measures (Individuals); *Sex Bias; *Sex Differences; *Test Construction; *Testing Problems

ABSTRACT

Three aspects of sex bias in ability tests are discussed. The first two areas of concern are the surface qualities of test items (surface, in the sense that these qualities may be seen by inspection and do not require empirical data), and the equating of male and female score distributions through item analysis and item selection techniques. The greater part of this paper focuses on any sex differences in test scores that remain after a test's content has been set, and how these differences may be handled to reduce or eliminate sex bias at the point when test interpretation leads to actions and plans for the future. The surface qualities of items include any aspect of their wording or their illustrations that suggest sexual inequality. "Sexist" language and art work, either within single items or summed across an entire test, are relatively easy to detect and eliminate through careful editing. (Author/GK)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

D.O. Herman

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

ED208057

Reducing Sex Bias in Ability Tests[1]

David O. Herman
Measurement Division
The Psychological Corporation
New York, New York

This paper will discuss three aspects of sex bias in ability tests. The first two areas of concern here are the surface qualities of test items (surface, in the sense that these qualities may be seen by inspection and do not require empirical data), and the equating of male and female score distributions through item analysis and item selection techniques. The greater part of this paper, however, will focus on any sex differences in test scores that remain after a test's content has been set, and how these differences may be handled to reduce or eliminate sex bias at the point when test interpretation leads to actions and plans for the future.

The surface qualities of items include any aspect of their wording or their illustrations that suggest sexual inequality. "Sexist" language and art work, either within single items or summed across an entire test, are relatively easy to detect and eliminate through careful editing. Guidelines for avoiding sexist language have been issued by a number of publishers, including the American Psychological Association; while their focus is primarily on the language of textbooks and scholarly materials, the principles also apply to test items. In any case, it is my impression that offensive, sexist items are seldom found on tests developed within the last few years.

Developing test items and tasks that appear sex-fair on the surface is only a start, however. Other things being equal, and they often are not, it is desirable to produce ability tests that minimize sex differences in raw scores. For one thing, if males and females perform about equally well on a test, then all discussion of how

TM 810 820

[1]Paper presented at the annual convention of the American Psychological Association, Los Angeles, August 24, 1981.

to handle sex differences when putting the scores to use is made unnecessary. But good practices in test construction can go only so far to produce instruments on which males and females do equally well. For instance, the items on some tests are highly homogeneous in nature and sex differences found on one item will tend to be found on the rest. I am thinking here of tests like Block Design on the Wechsler Preschool and Primary Scale of Intelligence, or measures of perceptual speed like the Minnesota Clerical Test. In addition, perhaps good practices in test construction should be less than fully effective at reducing sex differences on tests -- for example if the criteria by which test validity is judged are themselves sex-related. But this point probably warrants detailed study in its own right, and is only mentioned here in passing.

The Mechanical Reasoning tests of the Differential Aptitude Tests (DAT) illustrates a different kind of problem. Here the test items are intended to tap students' understanding of the principles of the physics of everyday life or, as one of the authors expressed it, "barnyard physics." Over the years it has been found that nearly all of the items so far tried out are easier for males than for females, though the size of these differences in difficulty varies considerably from item to item. Efforts have been made to minimize sex differences in total scores by choosing items with relatively small sex differences in difficulty, but these efforts must be subsidiary to other concerns such as including an adequate range of content coverage, selecting items with high discrimination values for females as well as males, and maintaining a suitable range of item difficulty for the test as a whole. So far this approach has had little effect on differences between the average scores of males and females.

From time to time critics have proposed that stylistic elements in the Mechanical Reasoning items bias the test toward males, and that items of another type would be preferable. The reasoning goes that if the item contexts could be made either totally

abstract or else more familiar to females, sex differences should be reduced. Unhappily, efforts to date to draft such items have not been very successful. Of course we will continue trying in the future, but for the present the Mechanical Reasoning test is regularly associated with sex differences in total raw scores.

The point here is that sex differences are seen on certain types of ability tests despite systematic attempts to eliminate them. In such cases one must then decide how best to compensate for these differences in such a way as to minimize adverse impact on persons of either sex when scores are interpreted and the consequences experienced. In general two quite different strategies have been used to achieve this end. One is to present separate norms for members of each sex, with the result that distributions of converted scores (percentiles, standard scores, etc.) will be the same for females and males, regardless of sex differences found at the raw-score level. The other strategy is to present norms based on combined-sex groups, but to warn users that males and females tend to perform differently on the test, and to suggest ways to compensate for these differences. The need for such warnings stems from the fact that combined-sex norms reproduce any sex differences that exist in the raw scores from which they are derived. In the case of the DAT, typical profiles based on combined-sex norms would show important sex differences, with average females scoring relatively high on Clerical Speed and Accuracy, Spelling, and Language Usage, and average males scoring high on Mechanical Reasoning and Space Relations. As a result, counselors operating from combined-sex percentiles on the DAT would tend to steer female counselees into occupations associated with clerical and language abilities, and male counselees into occupations associated with mechanical and spatial abilities -- in other words, into occupations that are stereotypic of their sex. The warnings mentioned above are needed to offset this sex-biased outcome.

Each strategy has its own advantages and disadvantages. Separate-sex norms assure roughly identical score distributions for males and females, and therefore will tend to produce equivalent counseling outcomes regardless of a counselor's sensitivity to matters of sex equity. At the same time, separate-sex norms require students to indicate their sex on their answer sheets so that the proper norms table may be used, thus calling students' attention to gender as a factor in the testing process, and raising the possibility of recording percentiles incorrectly if students have improperly coded their sex on the answer sheets. Further, using separate-sex norms may increase the possibility of clerical error in moving from raw scores to various converted scores, especially if the conversions are carried out by hand.

Combined-sex norms, too, have both desirable and undesirable effects. If norms are not differentiated by sex, examinees need not mark their sex on answer sheets for scoring purposes, with the result that scoring errors from this source are eliminated. However, scores of this type must be accompanied by suitable cautions about the existence of sex differences in the scores, and suggestions of how to allow for these differences in the interests of sex equity.

Which approach, then, is more likely to foster non-sexist score interpretation, sex-fair career guidance, and the like? The position taken here is that neither strategy is preferable for all tests in all situations. What are the factors in the testing process that should guide one's choice?

The suggestions that follow, by the way, are not specific to aptitude tests. Rather, they apply to any testing procedure whose purpose relates to career guidance, which would include vocational interest inventories and possibly personality inventories as well as aptitude and ability measures. It is generally accepted that within the context of guidance, the entire complex of testing and interpreting practices should have "exploration validity"; among other things this implies that

counselees of both sexes should be encouraged to consider the full range of career alternatives, including those traditionally filled by the opposite sex.

Let us assume that an aptitude or ability test is associated with sex differences and that the test developers have decided that some provision must be made to encourage sex-fair use of the sex-restrictive scores. One factor here is the relative ease of use of the two methods. Other things being equal, combined-sex norms are more convenient in that users need to juggle only one instead of two sets of norms, and that published norms tables will occupy half the space -- an increasingly significant point in these days of exploding paper and printing costs.

Another factor is the number of scores on the battery or inventory that show significant sex differences. If relatively few scores are associated with important sex differences, simple warnings to the user about the need to compensate for them may be sufficient. However, if males and females differ significantly on a large number of test scores, sex bias is difficult to handle with mere warnings, even to the most conscientious counselors. This will be especially true if males score higher on some scales and females higher on others. The sheer mental effort, not to mention memory, required to control for the sex differences is more than one should ask of a counselor, and here separate-sex norms seem clearly preferable.

Next, consider the size of the sex differences found on the scores to be reported. In general, the larger the sex differences, the more difficult it will be to interpret the scores in a sex-fair manner assisted only with a caution that users should keep these differences in mind when translating scores into advice to counselees, assessment of clients, and so forth.

A final issue: Do the sex differences vary systematically with age or grade level, or with any other variable by which the norms are presented? For example,

norms for the DAT are presented separately for each of five grades; for some of the scores, sex differences are fairly small for youngsters in grade 8, but become appreciable at the higher grade levels. It is far easier to compensate for such variable sex differences by using separate norms for males and females than with combined-sex norms and warnings about proper sex-fair usage.

A few examples of how these matters relate to actual testing practices may be in order at this point. The DAT has already been mentioned as a battery for which norms are presented separately by sex. The reasons are that five of its eight tests are associated with sex differences, three of which favor females and two of which favor males; that the size of some of these sex differences exceeds half a standard deviation at the mean; and that some of the differences tend to be larger in the higher than in the lower grades for which the battery is designed.

In the case of the Scholastic Aptitude Test (SAT) of the College Board, raw scores are transformed to the same set of standard scores regardless of the examinee's gender. These are the scores that are reported to examinees and colleges, and by their nature will retain any sex differences that exist at the raw-score level. The College Board does offer some interpretive materials that present percentile equivalents of the standard scores by sex, which may provide students and colleges with useful information, especially in the case of the SAT Mathematics score on which males as a group outperform females. Thus the primary reporting system for the SAT (scaled scores) yields generally lower scores for women on one of the two major areas, and may have adverse impact on women depending on how admissions committees use the test information. The supplementary separate-sex percentiles should help offset this possibility.

The several Wechsler intelligence scales consist of from 10 to 12 tests, scores on which combine to yield a Verbal, a Performance, and Full Scale IQ. It is a typical finding that the three IQs are associated with trivial sex differences, but that interesting sex differences may be found on the separate tests that make up the IQs. Only combined-sex norms are presented for the IQs because sex differences are so small here thay they may be disregarded. Combined-sex norms are also used for transforming raw scores to scaled scores on the separate tests, since even here the sex differences are in most cases not very large (though they may be statistically significant because of the large sample sizes). It is not at all certain that profile interpretation would benefit, or have fairer outcomes, if the test scores were evaluated in relation to a subject's own sex.

These few examples of current practices are hardly exhaustive, but should illustrate that different solutions may be effective in different situations. As for the future, there is a real possibility that sex differences in measured aptitudes will grow smaller over time. Continued efforts at innovative item writing and sensitive item selection may help reduce sex differences in total test scores while leaving the underlying test constructs unchanged. Furthermore, changes in the experiences typical of early childhood may result in boys and girls becoming more nearly like each other with respect to interests and activities, which may in turn influence young people's developed abilities. In this connection it is of interest that the British norms for the DAT show smaller sex differences than do the U. S. norms, a state of affairs that could bear study.

There is, however, no real reason to assume that the "true" distributions of all abilities are alike for males and females. There is some evidence, for instance, that sex differences in spatial ability may have a genetic component.

Study in this area is of considerable theoretical importance, for it may shed light on the processes by which spatial and other abilities develop. But in the context of this particular discussion, the genesis and modifiability of sex differences in ability are beside the point. At present the sexes differ significantly on some important ability dimensions and some action must be taken to ensure that these differences are not reflected in different and restricted sets of educational and career paths being suggested to human beings as a function of their sex. The choice of which action to take is not always clearcut, but will depend on a number of factors, some of which have been proposed here.